

**ROLL NO: 210701268**

**EXP 4: Create UDF in PIG**

**Step-by-step installation of Apache Pig on Hadoop cluster on Ubuntu Pre-requisite:**

- Ubuntu 16.04 or higher version running (I have installed Ubuntu on Oracle VM (Virtual Machine) VirtualBox),
- Run Hadoop on ubuntu (I have installed Hadoop 3.2.1 on Ubuntu 16.04). You may refer to my blog “How to install Hadoop installation” click [here](#) for Hadoop installation).

**Pig installation steps**

**Step 1:** Login into Ubuntu

**Step 2:** Go to <https://pig.apache.org/releases.html> and copy the path of the latest version of pig that you want to install. Run the following command to download Apache Pig in Ubuntu:

```
$ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz
```

**Step 3:** To untar pig-0.16.0.tar.gz file run the following command:

```
$ tar xvzf pig-0.16.0.tar.gz
```

**Step 4:** To create a pig folder and move pig-0.16.0 to the pig folder, execute the following command:

```
$ sudo mv /home/hadoop/pig-0.16.0 /home/hadoop/pig
```

**Step 5:** Now open the .bashrc file to edit the path and variables/settings for pig. Run the following command:

```
$ sudo nano .bashrc
```

Add the below given to .bashrc file at the end and save the file.

```
#PIG settings
export PIG_HOME=/home/hadoop/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/export
export PIG_CONF_DIR=$PIG_HOME/conf
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
#PIG setting
ends
```



The screenshot shows a terminal window titled "hadoop [Running] - Oracle VM VirtualBox". The terminal is running the GNU nano 2.5.3 editor, editing the file ".bashrc". The content of the file is as follows:

```
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"

export HIVE_HOME=/home/hadoop/apache-hive-3.1.2-bin
export PATH=$PATH:$HIVE_HOME/bin

#PIG settings
export PIG_HOME=/home/hadoop/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
```

The terminal window also shows a sidebar with various application icons and a bottom status bar with keyboard shortcuts like "Get Help", "Write Out", "Where Is", "Cut Text", "Justify", "Cur Pos", "Exit", "Read File", "Replace", "Uncut Text", "To Spell", and "Go To Line".

**Step 6:** Run the following command to make the changes effective in the .bashrc file:

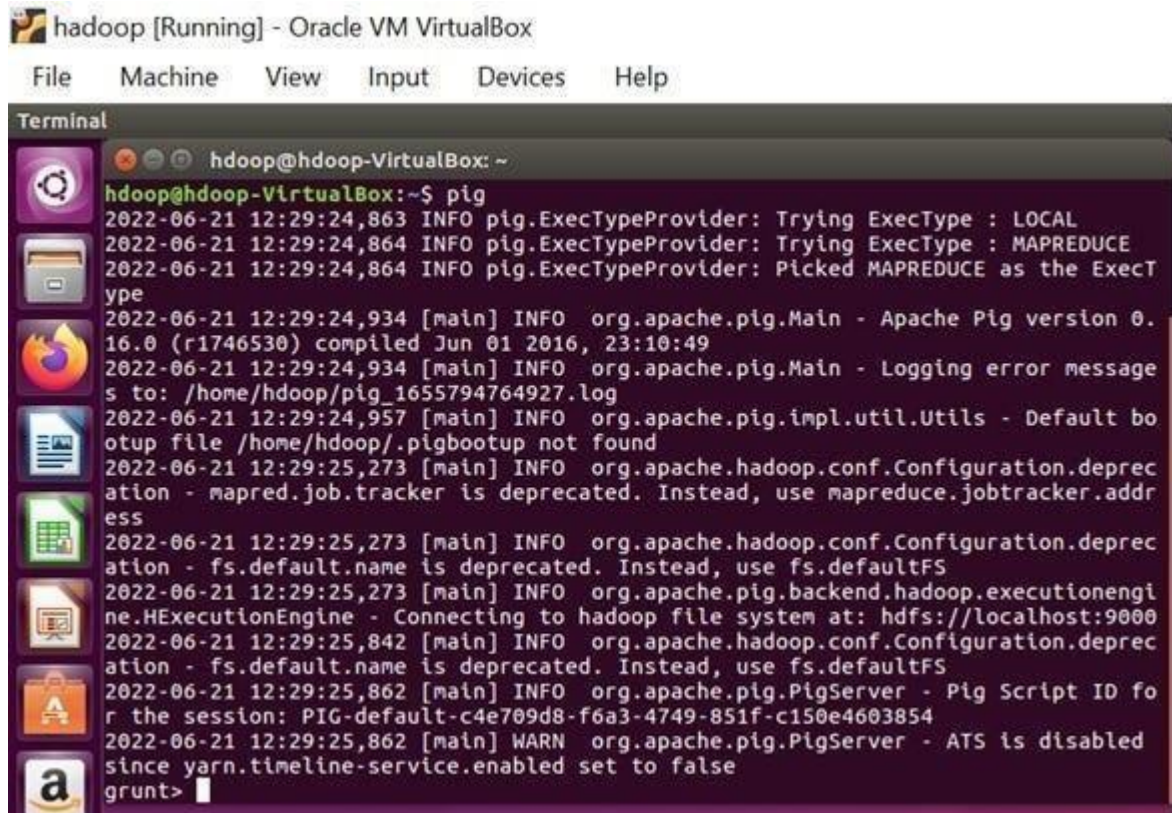
```
$ source .bashrc
```

**Step 7:** To start all Hadoop daemons, navigate to the hadoop-3.2.1/sbin folder and run the following commands:

```
$ ./start-dfs.sh$ ./start-yarn$ jps
```

**Step 8:** Now you can launch pig by executing the following command: \$

```
pig
```



```
hadoop [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Terminal
hadoop@hadoop-VirtualBox: ~
hadoop@hadoop-VirtualBox:~$ pig
2022-06-21 12:29:24,863 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2022-06-21 12:29:24,864 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2022-06-21 12:29:24,864 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecT
ype
2022-06-21 12:29:24,934 [main] INFO org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2022-06-21 12:29:24,934 [main] INFO org.apache.pig.Main - Logging error message
s to: /home/hadoop/pig_1655794764927.log
2022-06-21 12:29:24,957 [main] INFO org.apache.pig.impl.util.Utils - Default bo
otup file /home/hadoop/.pigbootup not found
2022-06-21 12:29:25,273 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-06-21 12:29:25,273 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-06-21 12:29:25,273 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2022-06-21 12:29:25,842 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-06-21 12:29:25,862 [main] INFO org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-c4e709d8-f6a3-4749-851f-c150e4603854
2022-06-21 12:29:25,862 [main] WARN org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt>
```

**Step 9:** Now you are in pig and can perform your desired tasks on pig. You can come out of the pig by the quit command:

> quit;

## CREATE USER DEFINED FUNCTION(UDF)

**Aim :** To create User Define Function in Apache Pig and execute it on map reduce.

### Procedure:

#### Create a sample text file

hadoop@Ubuntu:~/Documents\$ nano sample.txt

Paste the below content to sample.txt

- 1,John
- 2,Jane
- 3,Joe
- 4,Emma

hadoop@Ubuntu:~/Documents\$ hadoop fs -put sample.txt /home/hadoop/piginput/

---

**Create PIG File** hadoop@Ubuntu:~/Documents\$

nano demo\_pig.pig

**paste the below the content to demo\_pig.pig**

```
-- Load the data from HDFS  data = LOAD '/home/hadoop/piginput/sample.txt'
USING PigStorage(',') AS (id:int>
```

```
-- Dump the data to check if it was loaded correctly
DUMP data;
```

----- **Run**

**the above file** hadoop@Ubuntu:~/Documents\$ pig demo\_pig.pig

```
2024-08-07          12:13:08,791          [main]          INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil
- Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
```

-----

**Create udf file an save as uppercase\_udf.py**

uppercase\_udf.py

----- def

```
uppercase(text): return text.upper()
```

```
if __name__ == "__main__":
```

```
import sys for line in
```

```
sys.stdin:
```

```
line = line.strip() result
= uppercase(line)
print(result)
```

----- Create  
**the udfs folder on hadoop** `hadoop@Ubuntu:~/Documents$ hadoop fs -mkdir /home/hadoop/udfs`

**put the uppercase\_udf.py in to the abv folder** `hadoop@Ubuntu:~/Documents$ hdfs dfs -put uppercase_udf.py /home/hadoop/udfs/`

-----  
**hadoop@Ubuntu:~/Documents\$ nano udf\_example.pig** copy and  
**paste the below content on udf\_example.pig**

-- Register the Python UDF script

REGISTER 'hdfs:///home/hadoop/udfs/uppercase\_udf.py' USING jython AS udf;

-- Load some data data = LOAD 'hdfs:///home/hadoop/sample.txt'

AS (text:chararray);

-- Use the Python UDF   uppercased\_data = FOREACH data GENERATE  
udf.uppercase(text) AS uppercase\_text;

-- Store the result

STORE uppercased\_data INTO 'hdfs:///home/hadoop/pig\_output\_data';

-----  
**place sample.txt file on hadoop** `hadoop@Ubuntu:~/Documents$ hadoop fs -put sample.txt /home/hadoop/`

**To Run the pig file** `hadoop@Ubuntu:~/Documents$`

`pig -f udf_example.pig` **finally u get Success!**

**Job Stats (time in seconds):**

JobId Maps Reduces MaxMapTimeMinMapTime AvgMapTime MedianMapTime  
MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias  
Feature Outputs

job\_local1786848041\_0001 1 0 n/a n/a n/a n/a 00 0 0 data,uppercased\_data  
MAP\_ONLY hdfs:///home/hadoop/pig\_output\_data,

**Input(s):**

Successfully read 4 records (42778068 bytes) from: "hdfs:///home/hadoop/sample.txt"

**Output(s):**

Successfully stored 4 records (42777870 bytes) in: "hdfs:///home/hadoop/pig\_output\_data"

**Counters:**

Total records written : 4

Total bytes written : 42777870

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

**Job DAG:**

job\_local1786848041\_0001

2024-08-07 13:33:04,631 [main] WARN  
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker  
metrics system already initialized!

2024-08-07 13:33:04,639 [main] WARN  
org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker  
metrics system already initialized!

2024-08-07 13:33:04,644 [main] WARN

org.apache.hadoop.metrics2.impl.MetricsSystemImpl -  
JobTracker metrics system already initialized! 2024-08-07

13:33:04,667 [main] INFO

org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -  
Success!

**Note :**

**If any error check jython package is installed and check the path specified on the above steps are give correctly**

----- To  
**check the output file is created** hadoop@Ubuntu:~/Documents\$ hdfs dfs -ls  
/home/hadoop/pig\_output\_data

Found 2 items

If you need to examine the files in the output folder, use: To

**view the output** hadoop@Ubuntu:~/Documents\$ pig  
**demo\_pig.pig**

```
sudhashreem@sudhashreem-VirtualBox: ~/DA/exp4
6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:48,470 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:49,478 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:50,480 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:50,616 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-09-20 19:15:51,647 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:52,656 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:53,658 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:54,660 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:55,663 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:56,666 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:57,668 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:58,671 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:15:59,673 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:16:00,674 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried
9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 19:16:00,800 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retr
ieve job to compute warning aggregation.
2024-09-20 19:16:00,804 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-20 19:16:01,059 [main] INFO org.apache.pig.Main - Pig script completed in 4 minutes, 6 seconds and 308 milliseconds (246308
ms)
sudhashreem@sudhashreem-VirtualBox:~/DA/exp4$ hdfs dfs -cat /exp4/output/part-m-00000
1,JOHN
2,JANE
3,JOE
4,EMMA
sudhashreem@sudhashreem-VirtualBox:~/DA/exp4$
```

**Result:**

**Thus the program is executed successfully**

